同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

交通数据分析

第五讲 广义线性回归模型

沈煜 博士 副教授
嘉定校区交通运输工程学院311室
yshen@tongji.edu.cn
2022年03月25日

计划进度

周	日期	主讲	内容	模块
1	2022.02.25	沈煜	概述	
2	2022.03.04	沈煜	在线数据采集方法	爬虫
3	2022.03.11	沈煜	线性回归模型	回归分析
5	2022.03.18	沈煜	广义线性回归	
4	2022.03.25	沈煜	广义线性回归（作业1）	
6	2022.04.01	沈煜	空间数据描述性分析	
7	2022.04.08	沈煜	空间自回归方法（作业2）	
8	2022.04.15	沈煜	关联：Apriori	机器学习
9	2022.04.22	沈煜	决策树、支持向量机（作业3）	
10	2022.04.29	沈煜	浅层神经网络	
11	2022.05.06	沈煜	卷积神经网络（期末大作业）	
12	2022.05.13	沈煜	经典网络结构	
13	2022.05.20	沈煜	聚类：K-Means、DBSCAN	
14	2022.05.27	沈煜	贝叶斯方法、卡尔曼滤波	
15	2022.06.03	-	端午节放假	
16	2022.06.10	沈煜	期末汇报（1）	
17	2022.06.17	沈煜	期末汇报（2）	

主要内容

- 逻辑回归
- 计数回归

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

逻辑回归

Logistic Regression

广义线性回归

- y的分布服从某指数家族分布
- $p(y|x;\theta) = p(y|\eta) = b(y)\exp\left(\eta(\theta)T(y) - a(\eta)\right)$
- 一般$T(y) = y$（充分统计量sufficient statistic)
- 伯努利分布:
- 参数$\theta = \mu$
- $p(y;\mu) = \exp\left(\left(\log\left(\frac{\mu}{1-\mu}\right)\right)y + \log(1-\mu)\right)$
- 自然参数$\eta = \log\left(\frac{\mu}{1-\mu}\right) \Rightarrow \mu = \frac{1}{1+\exp(-\eta)}$
- $T(y) = y$
- $a(\eta) = -\log(1-\mu) = \log(1+\exp(\eta))$ (log-partition)
- $b(y) = 1$ (base measure)

逻辑回归模型

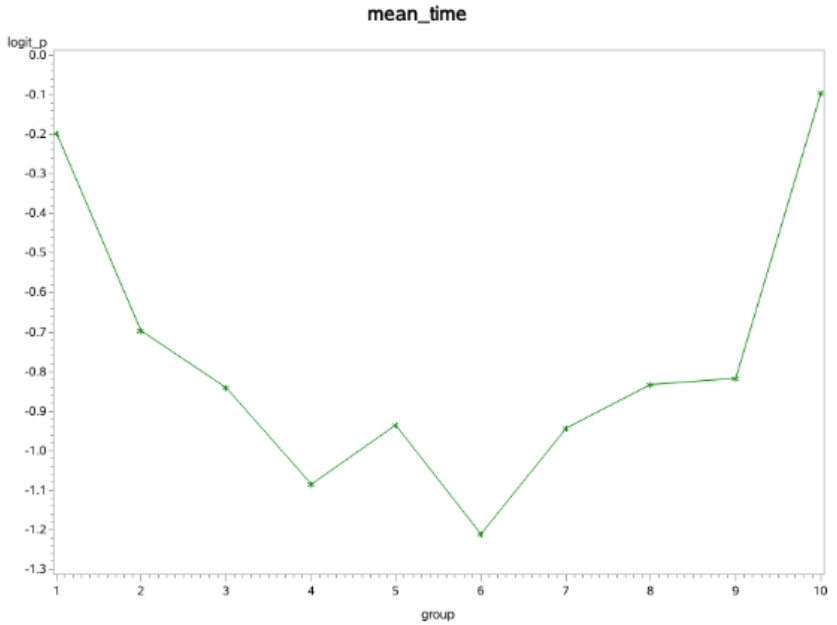
- 现象发生的概率（事故是否发生）
- $p = \frac{\exp(\beta_0+\beta X+\varepsilon)}{1+\exp(\beta_0+\beta X+\varepsilon)}$
- $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta X + \varepsilon$
- 比值/优势： $\text{Odds} = \frac{p}{1-p}$
- 比值比/优势比：$\text{OR} = \frac{\text{Odds}_1}{\text{Odds}_2}$
- $x = 1$相对$x = 0$，OR是$\exp(\beta)$倍

逻辑回归模型变量-选择

- 自变量与因变量的相关性分析
  - 卡方检验、T检验
- 进行自变量之间的相关性分析
  - 卡方检验、皮尔逊相关系数、T检验
- 进行多因素的逐步筛选
  - 逐步向前选择，逐步向后删除，向前向后结合

逻辑回归模型变量-形式

- 自变量与Logit(p)的关系并不是直线
- 如何判断？$\sim x^2$?$\sim |x-a|$?
- 将自变量按序排列
- 分成不同的组
- 统计每一组y=1的样本数和总样本数
- 计算每一组平均的p
- 计算logit（p）



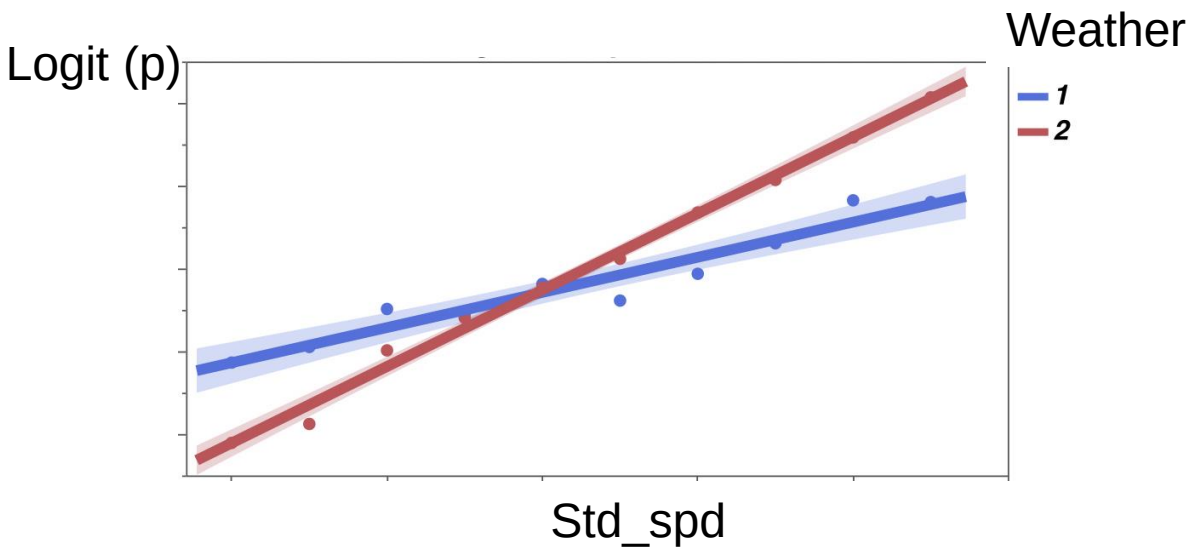
逻辑回归模型变量-形式

y	x1
1	1.1
0	1.5
1	1.2
0	2.6
1	2.1
0	2.5
0	2.7
0	3.2
1	3.5
1	3.8

- $x_1 \in (1,2), p = \frac{2}{3}$
- $logit(p) = \log\left(\frac{\frac{2}{3}}{1-\frac{2}{3}}\right) = \log 2$
- $x_1 \in (2,3), p = \frac{1}{4}$
- $logit(p) = \log\left(\frac{\frac{1}{4}}{1-\frac{1}{4}}\right) = -\log 3$
- $x_1 \in (3,4), p = \frac{2}{3}$
- $logit(p) = \log\left(\frac{\frac{2}{3}}{1-\frac{2}{3}}\right) = \log 2$

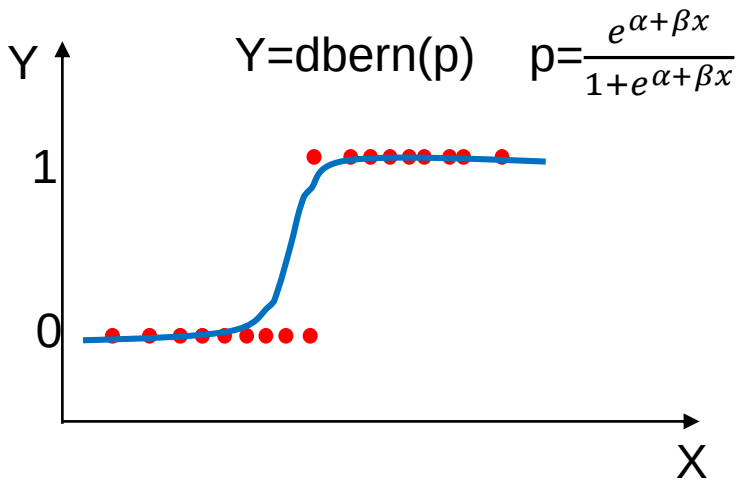
逻辑回归模型变量-交叉项

- 如果一个自变量对因变量的影响，随着另外一个自变量而发生变化，可以考虑加入交叉项
- $\mathrm{Logit(p)} = \ln\frac{p}{1-p} = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \beta_3 \mathrm{x}_1 \mathrm{x}_2$
  - 一个变量作为分类变量，另外一个变量分组，得到每组的Logit(p)，连起来；看分类变量的两条线是否相交
  - 一个变量作为分类变量，另外一个变量与logit（p）做回归



逻辑回归模型参数估计

- 逻辑回归模型是通过极大似然估计法得到
- 一个观测值的概率为
  - $P(y_i) = (p_i)^{y_i}(1-p_i)^{1-y_i}$
- 因各项观测相互独立，他们的联合分布可以表示为各观测值的乘积，也称为n个观测的似然函数
  - $L(\beta) = \prod_1^n (p_i)^{y_i}(1-p_i)^{1-y_i}$
  - $\ln(L(\beta)) = \sum_1^n [y_i \ln(p_i) + (1-y_i)\ln(1-p_i)]$



逻辑回归模型假设检验

- Wald检验，比较估计系数与0的差别

$$Z = \frac{\hat{\beta} - 0}{SE(\hat{\beta})}$$

- 似然比检验，通过对比两个相嵌套模型的对数似然函数统计量G

$$\mathrm{G} = \mathrm{G}_p - \mathrm{G}_k = -2\ln(L_p) + 2\ln(L_k)$$

  - 其中
  - G_p是$-2\ln(L_p)$，越小越能够代表拟合效果好
  - G_k是-2 $\ln(L_k)$
  - 模型p中的变量是模型k中变量的一部分，另一部分就是要检验的变量（模型p嵌套在模型k中）
  - G服从自由度为k-p的χ^2分布

逻辑回归模型拟合优度-分类表

- 分类表
  - 给定一个阈值，当p大于这个阈值时，y=1；否则y=0
  - 2×2的分类表，观测y vs 预测y

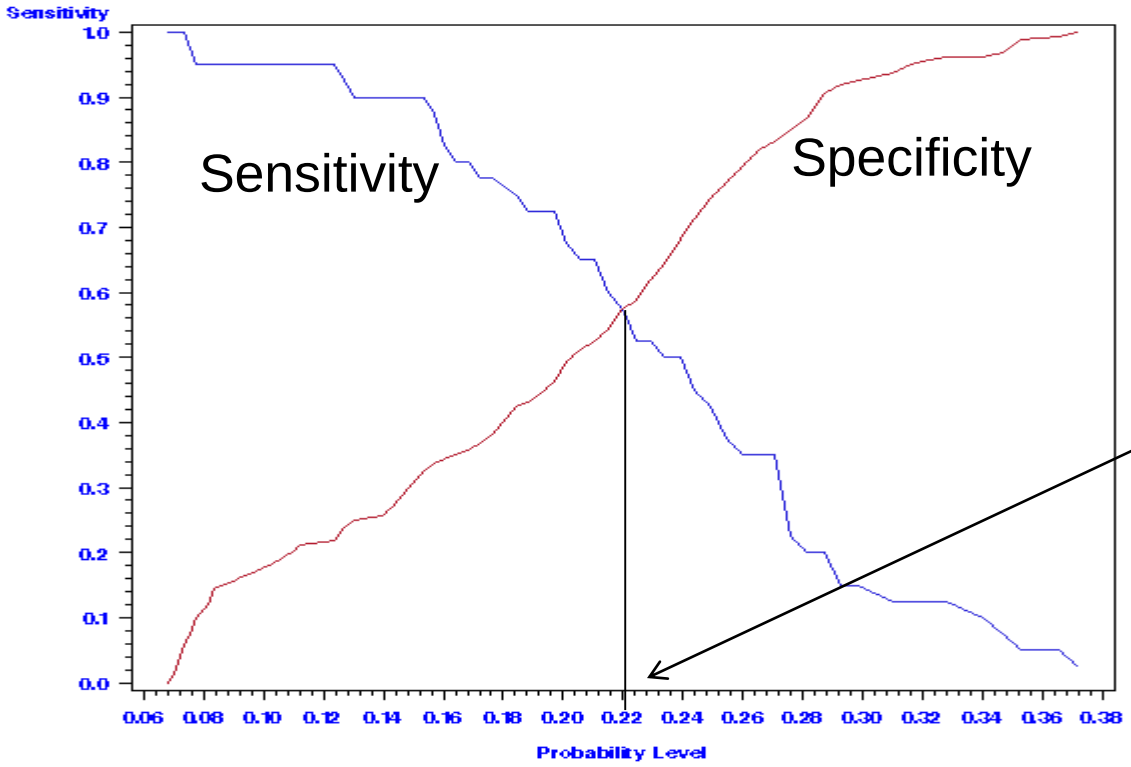
		预测y	
实际y		1	0
	1	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

  - 识别出的所有正例占所有正例的比例，敏感度（Sensitivity）=TP/(TP+FN)
  - 识别出的负例占所有负例的比例，特异度（Specificity）=TN/(FP+TN)
  - 将负例识别为正例的情况占所有负例的比例, False Positive Rate=FP/(FP+TN)
  - 将正例识别为负例的情况占所有正例的比例, False Negative Rate=FN/(TP+FN)

逻辑回归模型拟合优度-分类表

		预测y	
实际y		1	0
	1	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

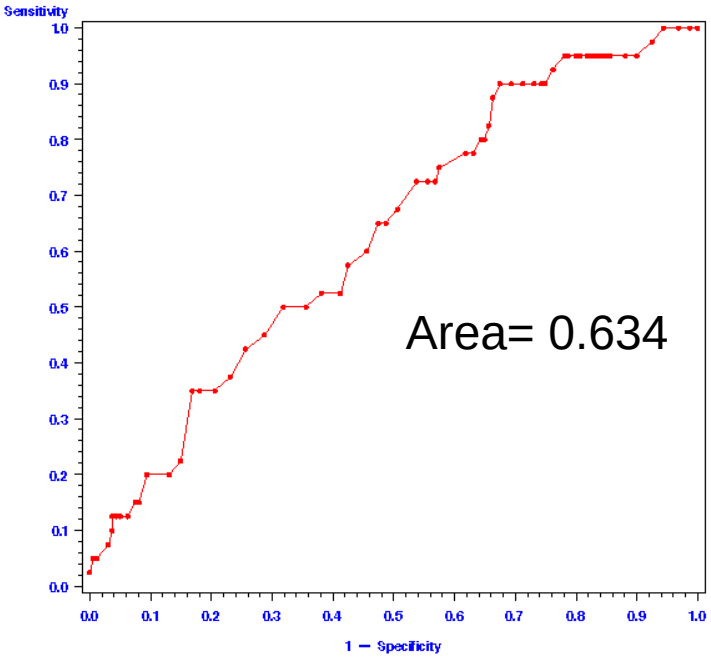
分类准确率=（TP+TN）/(TP+FN+FP+TN)



样本均衡时
最佳阈值

逻辑回归模型拟合优度-ROC

- 敏感度、特异度和分类准确率都依赖于阈值
- 提出来ROC（Receiver Operating Characteristic） Curve
  - 1. 画图法：针对所有的阈值敏感度vs特异度



面积反映模型能正确区分y=1和y=0的能力

ROC = 0.5: 没区分能力 (不比丢骰子好)
0.7 <= ROC < 0.8: 可接受
0.8 <= ROC < 0.9: 好
ROC > 0.9：非常好

逻辑回归模型拟合优度-ROC

- 2. 数值计算
  - 将实际y=1和y=0的两组数据分开，分别为数据组A（m样本量）和数据组B（n样本量）
  - 分别用模型计算两组数据中每一个样本y=1的概率
  - 将所有A和B的每一个样本进行对比，有m*n中组合
  - 每一个组合中，若数据组A样本y=1的概率大于数据组B样本y=1的概率，则改组合得分为1；若相等，得分为0.5；若小于得分为0
  - 将所有组的得分相加，得到数值q
  - ROC=q/mn

逻辑回归模型拟合优度-ROC

- 2. 数值计算
  - 举例

y=1	0.5
y=1	0.6
y=0	0.2
y=0	0.55
y=0	0.5

0.5	0.2	1
0.5	0.55	0
0.5	0.5	0.5
0.6	0.2	1
0.6	0.55	1
0.6	0.5	1

ROC=0.75

模型拟合优度对比

- 平均绝对偏差(Mean Absolute Deviation, MAD)

$$\frac{1}{m}\sum_{i=1}^{m}\left|(y_i-\hat{y}_i)\right|$$

- 均方误差（Mean Square Error, MSE)

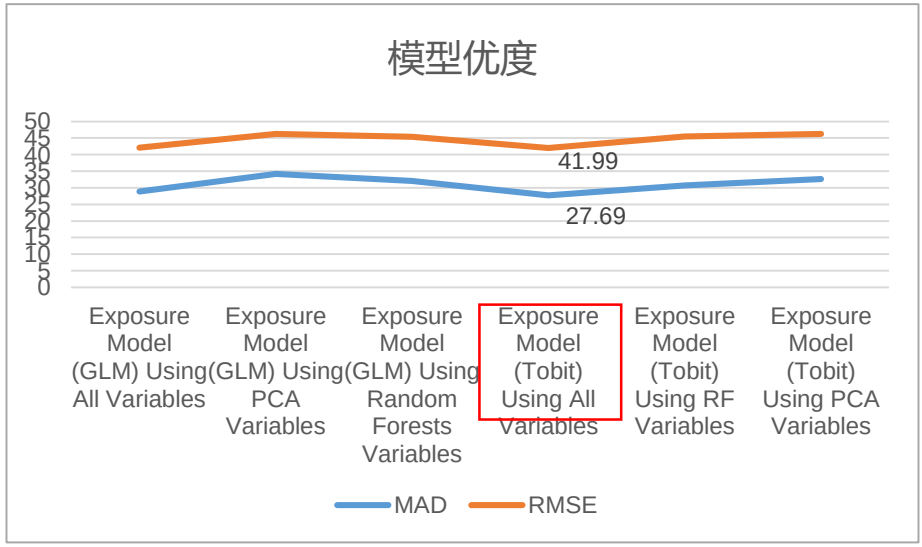
$$\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$$

- 均方根误差(Root Mean Square Error, RMSE)

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}$$

Table 2. Comparison of the exposure models developed.

Model Type (Exposure)	MAD	RMSE
Exposure Model (GLM) Using All Variables	28.91	42.12
Exposure Model (GLM) Using PCA Variables	34.21	46.26
Exposure Model (GLM) Using Random Forests Variables	32.06	45.40
Exposure Model (Tobit) Using All Variables	**27.69**	**41.99**
Exposure Model (Tobit) Using RF Variables	30.76	45.43
Exposure Model (Tobit) Using PCA Variables	32.61	46.24



模型拟合优度对比

- Akaike information criterion （AIC准则）
  - AIC=2k-2ln(L)
  - k是参数的数量, L是似然函数
- Bayesian information criterion（BIC准则）
  - BIC=ln(n)*k-2 ln(L)
  - n是样本量

不同的样本，是不能对比得出哪个模型更好的

Table 4. Three-year crash sum model standard error ratios with respect to random effects Poisson gamma model

Explanatory variable	Total	Rear-end	Sideswipe	Fixed object	All-other
LnAADT	1.023	0.998	1.026	1.013	1.005
LnLength	0.977	0.937	1.001	1.001	1.000
Urban rural indicator, 1 if rural, 0 if urban	0.983	0.967	1.006	1.002	–
Proportion of three or more lanes cross section	0.976	0.950	1.002	1.001	1.000
Number of horizontal curves per segment	0.973	0.941	1.002	1.000	–
Diamond interchange type indicator	0.983	0.966	1.004	1.001	–
Smallest vertical curve gradient in segment	0.978	–	1.001	–	–
Largest beginning vertical curve elevation	–	–	1.002	–	–
Longest horizontal curve central angle	–	–	–	1.000	–
Number of vertical curves	–	–	–	1.000	–
Shortest vertical curve length	0.976	–	–	–	–
Largest vertical curve rate of vertical curvature	0.968	–	–	–	–
Constant	1.022	0.996	1.026	1.013	1.005
γ_i^{-1} (dispersion parameter)	0.925	0.814	1.001	0.998	0.995
Log-likelihood at convergence	−1105.8	−669.8	−455.5	−493.9	−449.6
AIC	2233.6	1355.6	931.0	1007.7	909.1
BIC	2273.3	1384.5	967.1	1043.8	927.2
Sample size	274	274	274	274	274

Ratio = standard error of crash sum model/standard error of random effect Poisson gamma model.

案例数据

- **给定一组事故数据**
  - No. obs: 795
  - Case: 1(事故），0（非事故）
  - Weather: 1（下雨），0（晴天）
  - spd_dif_1min: 事故发生前0-1分钟内的速度变化值
  - spd_dif_2min: 事故发生前1-2分钟内的速度变化值
  - spd_dif_3min: 事故发生前2-3分钟内的速度变化值
  - spd_dif_4min: 事故发生前3-4分钟内的速度变化值
  - vol_dif_1min: 事故发生前0-1分钟内的流量变化值
  - vol_dif_2min: 事故发生前1-2分钟内的流量变化值
  - vol_dif_3min: 事故发生前2-3分钟内的流量变化值
  - vol_dif_4min: 事故发生前3-4分钟内的流量变化值

case	spd_dif_1min	spd_dif_2min	spd_dif_3min	spd_dif_4min	vol_dif_1min	vol_dif_2min	vol_dif_3min	vol_dif_4min
1	2.370056744	2.213838057	0.911692765	0.801681326	11.50133333	3.875	5.233333333	10.9375
0	2.233318186	2.826466355	0.500803885	0.550455975	7.614666667	7.392592593	7.237037037	2.614814815
0	0.409011611	2.027947031	1.011126828	1.123467889	7.488888889	11.42592593	8.696296296	1.62962963
0	0.5424272	1.393480201	1.287873812	0.03416836	5.885185185	1.666666667	5.908333333	6.791666667
0	2.974580882	2.681897594	1.014149125	1.353325272	7.368	2.297916667	2.085416667	10.19166667
1	2.858422352	1.449428804	1.221667064	2.261603667	4.512962963	5.16122449	2.87630854	3.697981771
0	2.355111004	1.529708825	5.778963516	0.995358896	0.408333333	1.388888889	2	5.266666667
0	0.989062788	2.83645343	4.03456893	0.477193624	2.939583333	2.2625	2.392592593	9.285185185
0	1.613592778	2.244937893	3.271009304	0	2.81875	2.583333333	2.62962963	0
0	0.125956284	1.949184934	2.310511822	1.028534169	2.5	1.25	1.781481481	3.516666667
1	2.532193025	1.29998667	0.860974691	2.399647986	10.307	6.27768595	11.45733333	7.728395062

案例要求

- 将Case作为因变量，其他的变量作为自变量，建立回归模型
  - 将样本分为70:30，模型标定：模型检验
  - 基于模型标定样本集
  - 进行建模准备
  - 模型构建与标定
  - 判断标定模型的拟合优度
  - 基于标定的模型，计算得到模型检验样本集的拟合优度

案例分析

- 分析变量性质
- 因变量：case（是否发生事故）
  - 二元变量
  - Logistic模型
- 自变量：天气、速度变化、流量变化
  - 相关性分析：没有冗余变量
  - 自变量筛选：变量形式、有无必要纳入交互项
- 回归模型检验
  - 假设检验
  - 拟合度及校验

案例分析：建模

- 样本随机分配为7比3

- **对所有自变量和因变量进行logistic建模**
- 如，使用R中的glm()函数
- 例：
  - logis_mod = glm(case ~ spd_dif_1min + spd_dif_2min + spd_dif_3min + spd_dif_4min + vol_dif_1min + vol_dif_2min + vol_dif_3min + vol_dif_4min + Weather, data_train, family = binomial(link = 'logit’))
  - logis_mod：包含训练集中所有自变量和因变量的logistic模型

案例分析：变量筛选

- **筛选变量**
- 如，R中的step()函数
  - 基于观测AIC与残差平方和最小的逐步回归函数
- 例：
  - step(object, scope, scale = 0, direction = c("both", "backward", "forward"), trace = 1, keep = NULL, steps = 1000, k = 2, …)
  - direction：逐步回归的方向，包括向前、向后以及两者结合的方法
  - 默认为向后回归，即将所有变量先放进模型中，逐步剔除不显著的变量，使得总体的 AIC最小

案例分析：变量筛选

- **筛选变量**
- logis_mod1 = step(logis_mod)
  - 对logis_mod进行逐步回归后的logistic模型

```
Start:  AIC=505.98
case ~ spd_dif_1min + spd_dif_2min + spd_dif_3min + spd_dif_4min +
    vol_dif_1min + vol_dif_2min + vol_dif_3min + vol_dif_4min +
    Weather

              Df Deviance    AIC
- Weather      1   486.10 504.10
- spd_dif_3min 1   486.26 504.26
- vol_dif_4min 1   486.32 504.32
- vol_dif_3min 1   486.94 504.94
- spd_dif_2min 1   487.35 505.35
<none>             485.98 505.98
- spd_dif_4min 1   490.02 508.02
- vol_dif_2min 1   490.17 508.17
- vol_dif_1min 1   493.12 511.12
- spd_dif_1min 1   497.96 515.96

Step:  AIC=504.1
case ~ spd_dif_1min + spd_dif_2min + spd_dif_3min + spd_dif_4min +
    vol_dif_1min + vol_dif_2min + vol_dif_3min + vol_dif_4min

              Df Deviance    AIC
- spd_dif_3min 1   486.36 502.36
- vol_dif_4min 1   486.39 502.39
- vol_dif_3min 1   487.09 503.09
- spd_dif_2min 1   487.50 503.50
<none>             486.10 504.10
- spd_dif_4min 1   490.10 506.10
- vol_dif_2min 1   490.35 506.35
- vol_dif_1min 1   493.33 509.33
- spd_dif_1min 1   498.32 514.32

Step:  AIC=502.36
case ~ spd_dif_1min + spd_dif_2min + spd_dif_4min + vol_dif_1min +
    vol_dif_2min + vol_dif_3min + vol_dif_4min

              Df Deviance    AIC
- vol_dif_4min 1   486.59 500.59
- vol_dif_3min 1   487.53 501.53
- spd_dif_2min 1   488.11 502.11
<none>             486.36 502.36
- vol_dif_2min 1   490.45 504.45
- spd_dif_4min 1   491.29 505.29
- vol_dif_1min 1   493.83 507.83
- spd_dif_1min 1   498.53 512.53
```

```
Step:  AIC=500.59
case ~ spd_dif_1min + spd_dif_2min + spd_dif_4min + vol_dif_1min +
    vol_dif_2min + vol_dif_3min

              Df Deviance    AIC
- vol_dif_3min 1   488.04 500.04
- spd_dif_2min 1   488.46 500.46
<none>             486.59 500.59
- vol_dif_2min 1   490.68 502.68
- spd_dif_4min 1   491.81 503.81
- vol_dif_1min 1   494.50 506.50
- spd_dif_1min 1   499.22 511.22

Step:  AIC=500.04
case ~ spd_dif_1min + spd_dif_2min + spd_dif_4min + vol_dif_1min +
    vol_dif_2min

              Df Deviance    AIC
- spd_dif_2min 1   489.79 499.79
<none>             488.04 500.04
- spd_dif_4min 1   492.98 502.98
- vol_dif_2min 1   493.54 503.54
- vol_dif_1min 1   496.17 506.17
- spd_dif_1min 1   501.35 511.35

Step:  AIC=499.79
case ~ spd_dif_1min + spd_dif_4min + vol_dif_1min + vol_dif_2min

              Df Deviance    AIC
<none>             489.79 499.79
- spd_dif_4min 1   495.62 503.62
- vol_dif_2min 1   496.38 504.38
- vol_dif_1min 1   497.76 505.76
- spd_dif_1min 1   505.48 513.48
```

在逐步筛选变量的过程中AIC逐渐减小
每一步都删除了一个变量使得 AIC减小
最后AIC值为 499.79

案例分析：参数分析

- 最后筛选出的变量
  - spd_dif_1min、spd_dif_4min、vol_dif_1min、vol_dif_2min
- 分析参数
  - 如R中通过summary()函数
  - 系数影响均显著

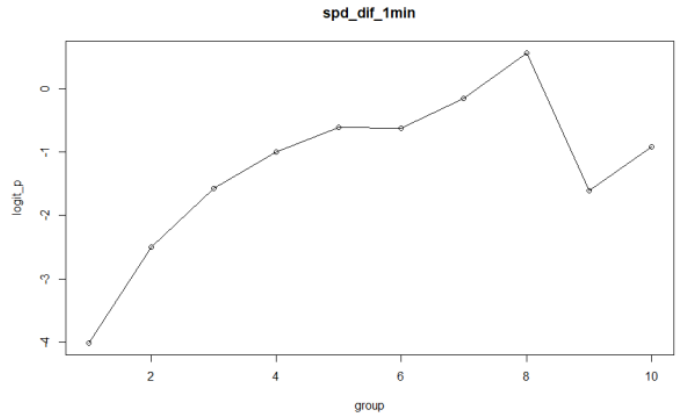
```
Call:
glm(formula = case ~ spd_dif_1min + spd_dif_4min + vol_dif_1min +
    vol_dif_2min, family = binomial(link = "logit"), data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6679  -0.5999  -0.5027  -0.4183   2.1724

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.69721    0.22908 -11.774  < 2e-16 ***
spd_dif_1min  0.13504    0.03348   4.034  5.5e-05 ***
spd_dif_4min  0.10154    0.04131   2.458  0.01398 *
vol_dif_1min  0.03848    0.01324   2.908  0.00364 **
vol_dif_2min  0.03880    0.01490   2.605  0.00919 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 541.72  on 555  degrees of freedom
Residual deviance: 489.79  on 551  degrees of freedom
AIC: 499.79

Number of Fisher Scoring iterations: 4
```

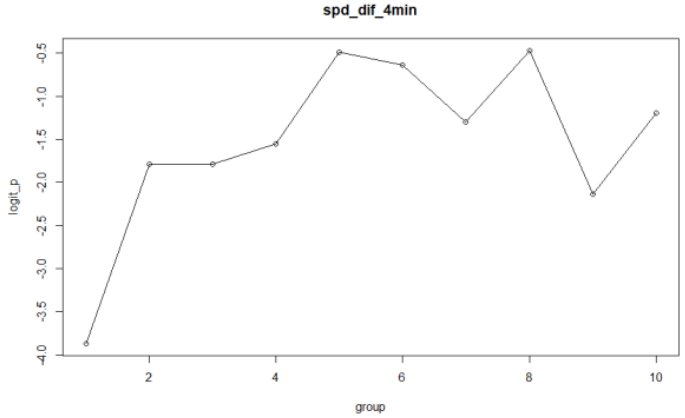
案例分析：变量形式转换

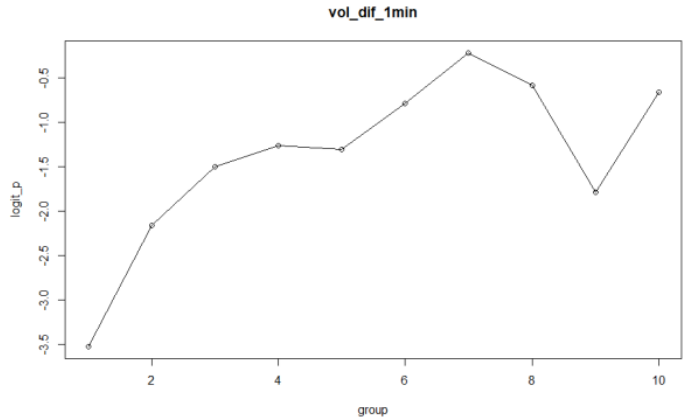
- 自变量与因变量有可能不是线性相关，但此变量有显著影响，因此需要进行一定的转换
- 对四个变量按顺序排列，对各变量数据分成不同等间距的组（如9分位数）
- 统计每组case=1的个数与样本总数比值p
- 计算logit(p)并按组顺序绘制图形

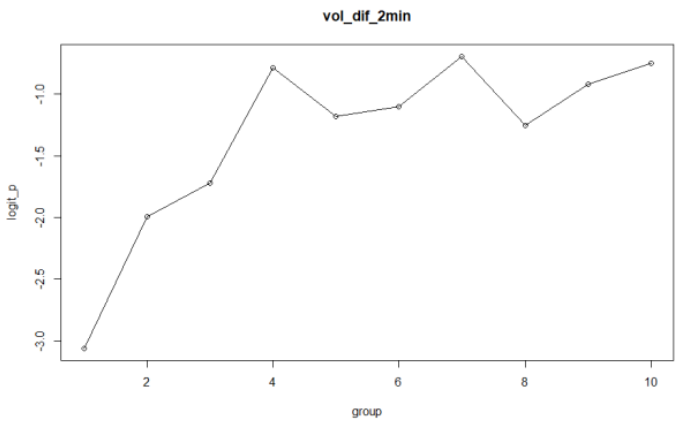
案例分析：变量形式转换

- 可以看出每个变量logit值和分组编号正相关
- 可以认为四个变量和因变量线性相关
- 不需要进行变量形式的转换









案例分析：交叉项判断

- 如果一个自变量对因变量的影响，随着另外一个自变量而发生变化，可以考虑加入交叉项
- 目前筛选出来的变量
  - spd_dif_1min：事故发生前1分钟内的速度变化值
  - spd_dif_4min：事故发生前4分钟内的速度变化值
  - vol_dif_1min：事故发生前1分钟内的流量变化值
  - vol_dif_2min：事故发生前 2分钟内的流量变化值
- 有可能交叉的：
  - spd_dif_1min*spd_dif_4min
  - vol_dif_1min*vol_dif_2min
  - spd_dif_1min*vol_dif_1min
- 进行三组交叉项实验

案例分析：交叉项判断

- 加入spd_dif_1min * spd_dif_4min建立logistic模型

```
Call:
glm(formula = case ~ spd_dif_1min + spd_dif_4min + vol_dif_1min +
    vol_dif_2min + spd_dif_1min * spd_dif_4min, family = binomial(link = "logit"),
    data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6648  -0.6017  -0.5023  -0.4174   2.1742

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.7063269  0.2582036 -10.481  < 2e-16 ***
spd_dif_1min               0.1374492  0.0458327   2.999  0.00271 **
spd_dif_4min               0.1048375  0.0595409   1.761  0.07828 .
vol_dif_1min               0.0384499  0.0132438   2.903  0.00369 **
vol_dif_2min               0.0388299  0.0149021   2.606  0.00917 **
spd_dif_1min:spd_dif_4min -0.0007863  0.0102264  -0.077  0.93871
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 541.72  on 555  degrees of freedom
Residual deviance: 489.78  on 550  degrees of freedom
AIC: 501.78

Number of Fisher Scoring iterations: 4
```

- AIC提升，但交叉项与spd_dif_4min的系数不显著
- 不考虑加入该交叉项

案例分析：交叉项判断

- 加入vol_dif_1min * vol_dif_2min建立logistic模型

```
Call:
glm(formula = case ~ spd_dif_1min + spd_dif_4min + vol_dif_1min +
    vol_dif_2min + vol_dif_1min * vol_dif_2min, family = binomial(link = "logit"),
    data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6221  -0.6024  -0.4995  -0.4108   2.1854

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.7658330  0.2551868 -10.838  < 2e-16 ***
spd_dif_1min             0.1342702  0.0335579   4.001  6.3e-05 ***
spd_dif_4min             0.1024323  0.0413276   2.479   0.0132 *
vol_dif_1min             0.0455534  0.0172841   2.636   0.0084 **
vol_dif_2min             0.0482458  0.0209704   2.301   0.0214 *
vol_dif_1min:vol_dif_2min -0.0007895  0.0012446  -0.634   0.5258
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 541.72  on 555  degrees of freedom
Residual deviance: 489.39  on 550  degrees of freedom
AIC: 501.39

Number of Fisher Scoring iterations: 4
```

- AIC提升，但交叉项系数不显著
- 不考虑加入该交叉项

案例分析：交叉项判断

- 加入vol_dif_1min * spd_dif_1min建立logistic模型

```
Call:
glm(formula = case ~ spd_dif_1min + spd_dif_4min + vol_dif_1min +
    vol_dif_2min + vol_dif_1min * spd_dif_1min, family = binomial(link = "logit"),
    data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7527  -0.6082  -0.5004  -0.4005   2.1931

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.850504   0.256809 -11.100  < 2e-16 ***
spd_dif_1min              0.191280   0.051795   3.693 0.000222 ***
spd_dif_4min              0.102789   0.041543   2.474 0.013350 *
vol_dif_1min              0.061883   0.021192   2.920 0.003499 **
vol_dif_2min              0.039177   0.014986   2.614 0.008943 **
spd_dif_1min:vol_dif_1min -0.007759   0.005488  -1.414 0.157415
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 541.72  on 555  degrees of freedom
Residual deviance: 487.74  on 550  degrees of freedom
AIC: 499.74

Number of Fisher Scoring iterations: 4
```

- AIC提升，但交叉项系数不显著
- 不考虑加入该交叉项

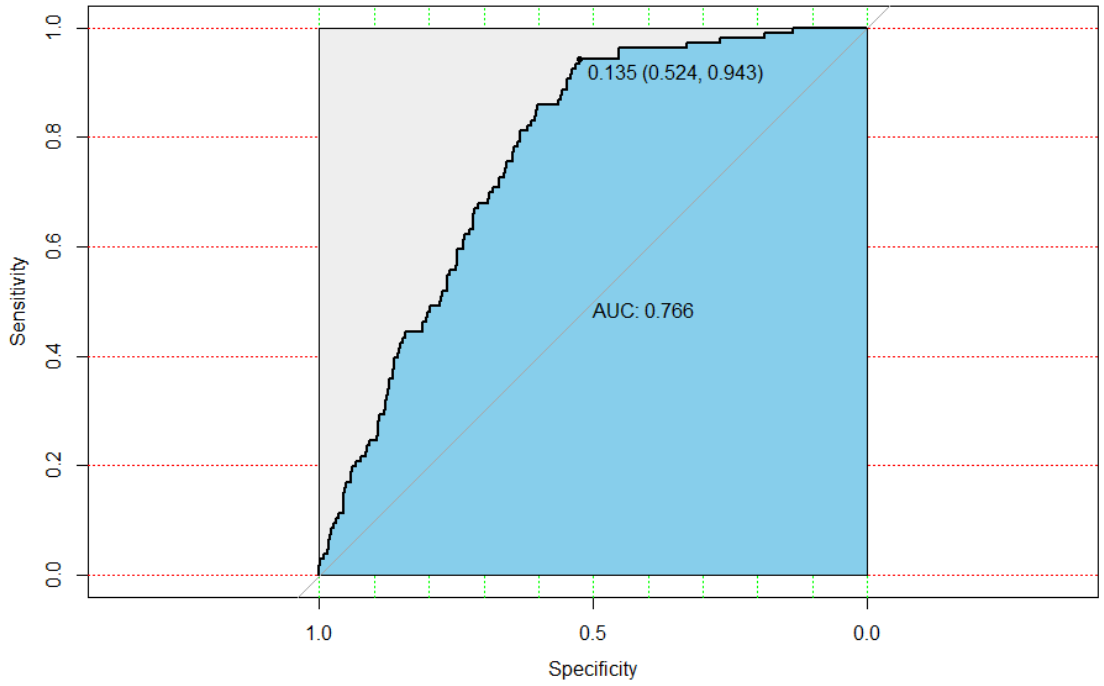
案例分析：假设检验

- 模型变量
  - 因变量：case
  - 自变量：spd_dif_1min、spd_dif_4min、vol_dif_1min、vol_dif_2min
- 分析结果：变量影响均显著

```
Call:
glm(formula = case ~ spd_dif_1min + spd_dif_4min + vol_dif_1min +
    vol_dif_2min, family = binomial(link = "logit"), data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6679  -0.5999  -0.5027  -0.4183   2.1724

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.69721    0.22908 -11.774  < 2e-16 ***
spd_dif_1min  0.13504    0.03348   4.034  5.5e-05 ***
spd_dif_4min  0.10154    0.04131   2.458  0.01398 *
vol_dif_1min  0.03848    0.01324   2.908  0.00364 **
vol_dif_2min  0.03880    0.01490   2.605  0.00919 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 541.72  on 555  degrees of freedom
Residual deviance: 489.79  on 551  degrees of freedom
AIC: 499.79

Number of Fisher Scoring iterations: 4
```

案例分析：拟合优度

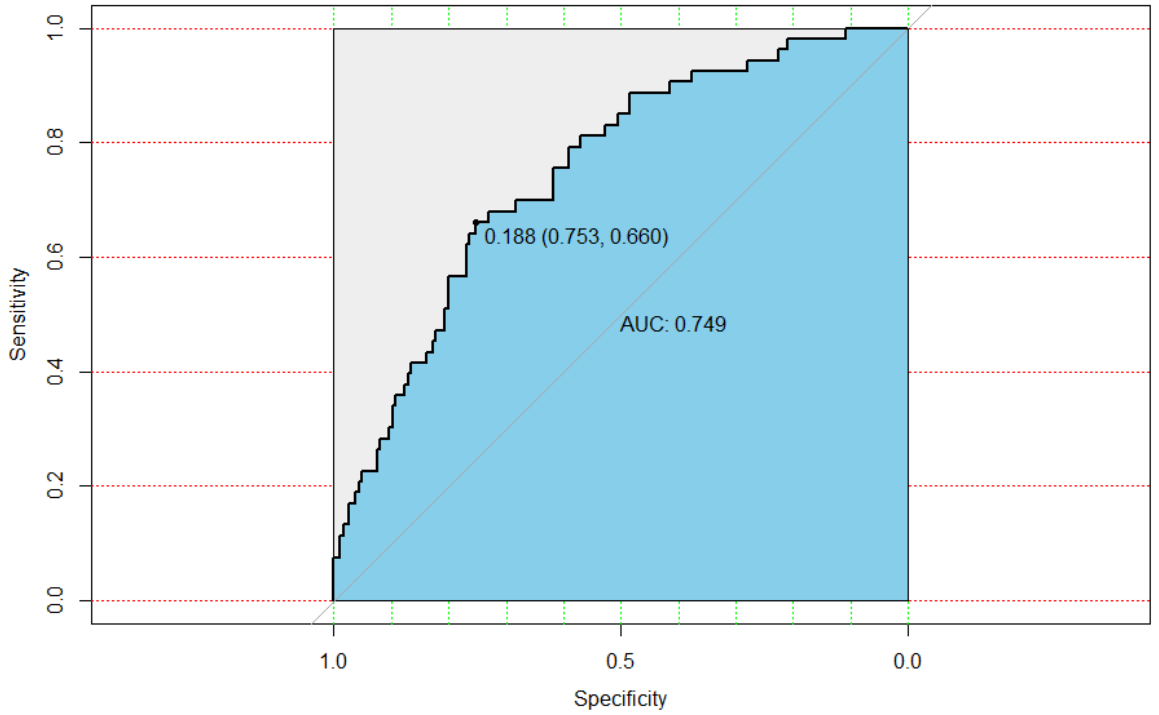
- 绘制ROC曲线计算 ROC值
  - ROC等于0.5：没有区分能力
  - 大于0.7小于0.8：可以接受
  - 大于 0.8小于 0.9：拟合优度好
  - 大于 0.9：非常好



ROC值为0.766
模型可以接受

案例分析：模型校验

- 用样本30%的数据作为测试集对模型进行拟合优度的判别



ROC值为0.749
与标定模型拟合优度相近
模型可以接受

作业

- 数据：运输经济学共享单车数据集
- 因变量为骑车与不骑车（不骑车为0，非0为骑车，转换为1）
- 要求：
- 1、将样本分为70:30，模型标定：模型检验
- 2、基于模型标定样本集
- 3、进行建模准备
  - 数据清洗等步骤参照之前上课内容
- 4、模型构建与标定（包括变量形式、交叉项等考虑）
- 5、判断标定模型的拟合优度
- 6、基于标定的模型，计算得到模型检验样本集的拟合优度

作业要求

- **提交时间：2022年4月8日上午9点59分**
- **提交邮箱：945441387@qq.com（吕叶婷）**

- **格式要求：**
- 不许超过12页
- 超一页扣一分
- 默认页边距、小四（12号）、1.5倍行距
- 宋体（英文、数字可以是Times New Roman或其他衬线字体）

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

计数数据回归

计数数据模型

- 计数数据模型
  - 因变量是计数数据（非负整数），给定的一段时间内，一个特定时间发生的次数
    - 一年内路段发生的事故数量
    - 一年内有多少辆车新注册
    - 一年内公交车的乘客量等

- 对于经典的线性模型：

$$Y=X\beta+\varepsilon$$

- 左端为非负整数；右端没有限制
- 采用对数变换， $\log(Y)=X\beta+\varepsilon$,解决非负限制问题
- 而且研究中发现挺多y有可能是0

计数数据模型-泊松回归模型

- 针对被解释变量观测值的非负整数特征
  - Gilbert (1979) 泊松回归模型 (Poisson Regression Model)
  - $f(y|\mu) = \frac{(\mu t)^{y} e^{-\mu t}}{y!}$
  - t是时间长度
  - μ单位时间内的期望
  - 泊松分布的重要特征是均值和方差相同，当t=1时
    - E（y）=μ,Var(y)=μ,称为分散均衡（equidispersion）
    - 意思是针对同一个个体，例如某条路段一年内发生事故的次数，进行无数次重复抽样，得到的计数数据序列的均值和方差相等
    - 在实际中，很难或者不可能重复抽样的，只能根据不同个体的一次抽样，计算多个个体的均值和方差

计数数据模型-负二项回归

- Hausman,Hall & Griliches(1984)负二项回归模型（Negative Binomial model-NB model）

- 二项分布（binomial distribution）
  - 重复n次独立的伯努利实验
  - 随机变量x服从参数为n和p的二项分布
  - X~B（n,p），0<p<1即
  - P(x=k)=$C_n^k p^k (1-p)^{n-k}$
  - 均值是E（x）=np，方差是Var（x）=np(1-p)
  - 计数过程中的均值大于方差，称为分散不足（underdispersion）

计数数据模型-负二项回归

- 满足以下条件的称为负二项分布
- ①实验包含一系列独立的实验， 每个实验都有成功、失败两种结果，成功的概率是恒定的，实验持续到r次成功,所需要的试验数X
  - P(x=k)=$C_{k-1}^{r-1}p^r(1-p)^{k-r}$
  - E(x)=$\frac{r}{p}$, Var(x)=$\frac{r(1-p)}{p^2}$
- ②实验包含一系列独立的实验， 每个实验都有成功、失败两种结果，成功的概率是恒定的,出现r次失败前成功的次数X
  - P(x=k,r,p)=$C_{k+r-1}^{\mathrm{k}}p^k(1-p)^r$
  - k次成功，r次不成功，成功概率为p
  - E(x)=$\frac{r(1-p)}{p}$, Var(x)=$\frac{r(1-p)}{p^2}$
  - 如果计数过程中的均值小于方差

    称为分散过度（overdispersion）

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

第五讲 结束